Microbial genome evolution - the pan-genome

Pablo Vinuesa

Centro de Ciencias Genómicas – UNAM

vinuesa @ ccg_unam_mx

http://www.ccg.unam.mx/~vinuesa/

@pvinmex

Introducción a la Filoinformática:

Pan-genómica y Filogenómica microbiana –

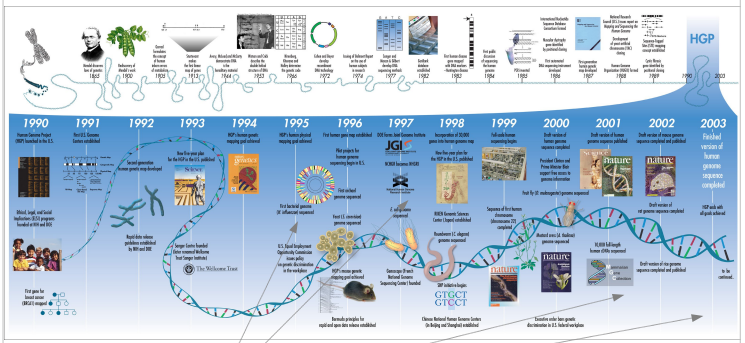
NNB & CCG-UNAM, 1-5 Agosto 2022

https://github.com/vinuesa/TIB-filoinfo

Genome evolution and the pan-genome

Contents

Genome sequencing and the pan-genome



1995, *H. influenzae* complete genome
1997, *E. coli* K12 (commensal)
2001, *E. coli* O157:H7 (EHEC)
2002, *E. coli* CFT073 (UPEC)

Sanger sequencing (dideoxy-chain termination) of reference genomes, mostly single individuals from different model species

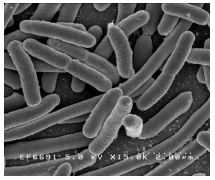
Basic concepts in bacterial comparative genomics - Mosaic genome structure and the pan-genome

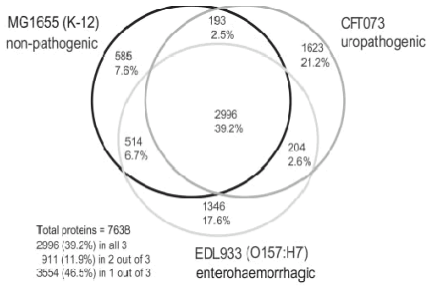
Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*

R. A. Welch*, V. Burland†‡, G. Plunkett III†, P. Redford*, P. Roesch*, D. Rasko§, E. L. Buckles¶, S.-R. Liou†, A. Boutin†**, J. Hackett*,††, D. Stroud†, G. F. Mayhew†, D. J. Rose†, S. Zhou†‡‡, D. C. Schwartz‡‡, N. T. Perna§§, H. L. T. Mobley§, M. S. Donnenberg¶, and F. R. Blattner†

17020–17024 | PNAS | December 24, 2002 | vol. 99 | no. 26

Edited by John J. Mekalanos, Harvard Medical School, Boston, MA, and approved October 24, 2002 (received for review August 30, 2002)





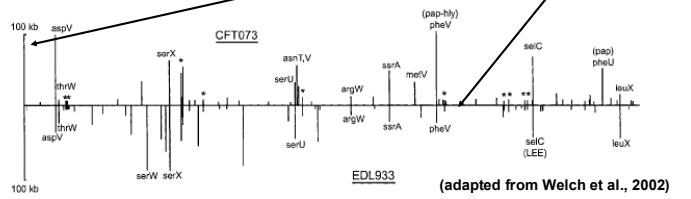
Basic concepts in bacterial comparative genomics – Mosaic genome structure and the pan-genome

"accessory/flexible genome (shell + cloud genes)":

- patchy distribution among strains
- ecological specialization
- Plasmids, phages, GIs, ICE, integrons, ISs …
- highly transferable (HGT)

"core genome":

- orthologous gene families
- species/population trees
- population genetics
- Genome annotation



(adapted from Welch et al., 2002)

Fig. 3. Locations and sizes of CFT073 and EDL933 islands. Island size, vertical axis; position in colinear backbone, horizontal axis. All islands >4 kb are shown. Islands located at tRNAs are indicated by tRNA labels. One tmRNA (*ssrA*) is also an insertion target. *, CFT073 and EDL933 islands in the same backbone location but not near tRNAs.

Basic concepts in bacterial comparative genomics – Mosaic genome structure and the pan-genome

PNAS

Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome"

Hervé Tettelin[a,b], Vega Masignani[b,c], Michael J. Cieslewicz[b,d,e], Claudio Donati[c], Duccio Medini[c], Naomi L. Ward[a,f], Samuel V. Angiuoli[a], Jonathan Crabtree[a], Amanda L. Jones[g], A. Scott Durkin[a], Robert T. DeBoy[a], Tanja M. Davidsen[a], Marirosa Mora[c], Maria Scarselli[c], Immaculada Margarit y Ros[c], Jeremy D. Peterson[a], Christopher R. Hauser[a], Jaideep P. Sundaram[a], William C. Nelson[a], Ramana Madupu[a], Lauren M. Brinkac[a], Robert J. Dodson[a], Mary J. Rosovitz[a], Steven A. Sullivan[a], Sean C. Daugherty[a], Daniel H. Haft[a], Jeremy Selengut[a], Michelle L. Gwinn[a], Liwei Zhou[a], Nikhat Zafar[a], Hoda Khouri[a], Diana Radune[a], George Dimitrov[a], Kisha Watkins[a], Kevin J. B. O'Connor[h], Shannon Smith[i], Teresa R. Utterback[i], Owen White[a], Craig E. Rubens[g], Guido Grandi[c], Lawrence C. Madoff[e,j], Dennis L. Kasper[e,j], John L. Telford[c], Michael R. Wessels[d,e], Rino Rappuoli[c,k,l], and Claire M. Fraser[a,b,k,m]

PNAS | September 27, **2005** | vol. 102 | no. 39 | 13951

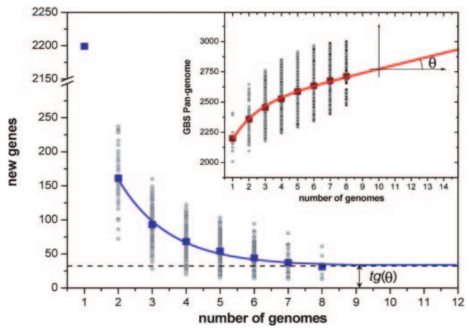
… Unfortunately, the sequence of a single genome does not reflect how genetic variability drives pathogenesis within a bacterial species and also limits genome-wide screens for vaccine candidates or for antimicrobial targets.

We have generated the genomic sequence of six strains representing the five major disease-causing serotypes of *Streptococcus agalactiae*, the main cause of neonatal infection in humans. Analysis of these genomes and those available in databases showed that the *S. agalactiae* species can be described by a **pan-genome** consisting of a **core genome** shared by all isolates, accounting for 80% of any single genome, plus a **dispensable genome** consisting of partially shared and strain-specific genes.

Mathematical extrapolation of the data suggests that the **gene reservoir** available for inclusion in the S. agalactiae **pan-genome is vast** and that unique genes will continue to be identified even after sequencing hundreds of genomes.

Basic concepts in bacterial comparative genomics – Mosaic genome structure and the pan-genome

The **core** and **pan-genome** for 8 *Streptococcus agalactiae* genomes



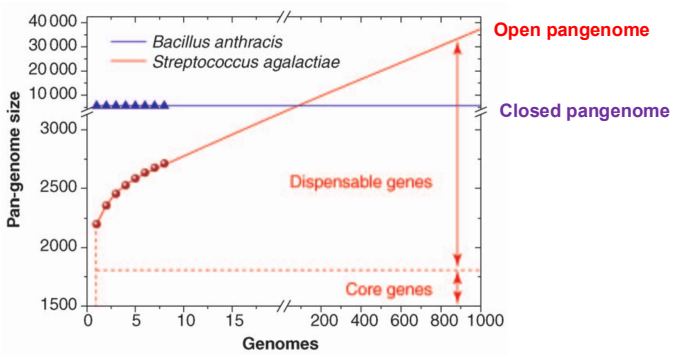
pangenome

core-genome

Species pan-genome = **core genome** shared by all isolates **+**
dispensable genome consisting of partially shared and strain-specific genes

Tettelin et al. (2005) PNAS | 102 | 39 | 13951

Basic concepts in bacterial comparative genomics – the pan-genome (Tettelin et al. 2005)



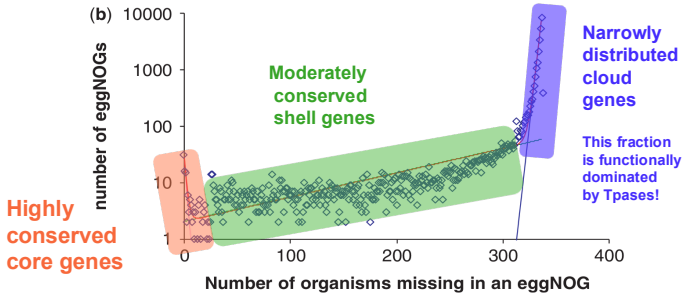
Open pangenome

Closed pangenome

Tettelin, H., et al. (2008). "Curr. Opin. Microbiol. **11(5): 472-7.**

Structure and the bacterial pan-genome

Distribuion of COGs by the number of organisms included in each cluster:

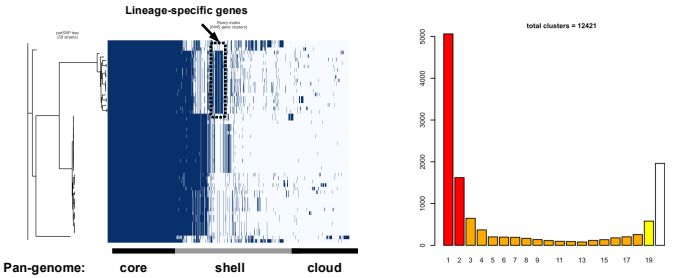


Fit of an **exponential decay function with three exponents**:

The core (~70 clusters), shell (~5700 clusters) the cloud (~ 24,000 clusters)

Koonin & Wolf, NAR 2008. Vol 36, No. 21

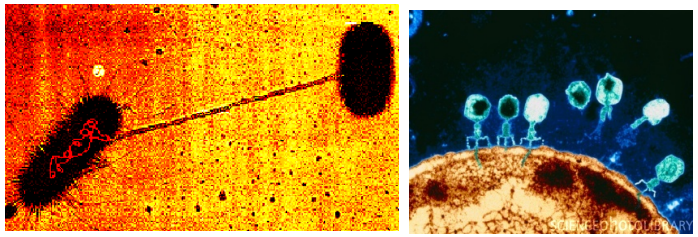
Structure of the bacterial pan-genome



Key questions in pan-genome research

- **Which are the sources and mechanisms of within-species genome content variability?**
- **How can pan-genomes be analyzed and modeled?**
- **What is the impact of population structure on pan-genome analysis and phylogenomic inference?**
- **Why do bacterial species have pan-genomes?**

Evolution of the pan-genome: prokaryotic sex (parasexuality) and HGT



- **Conjugation**
- **Transduction**
- **Transformation**

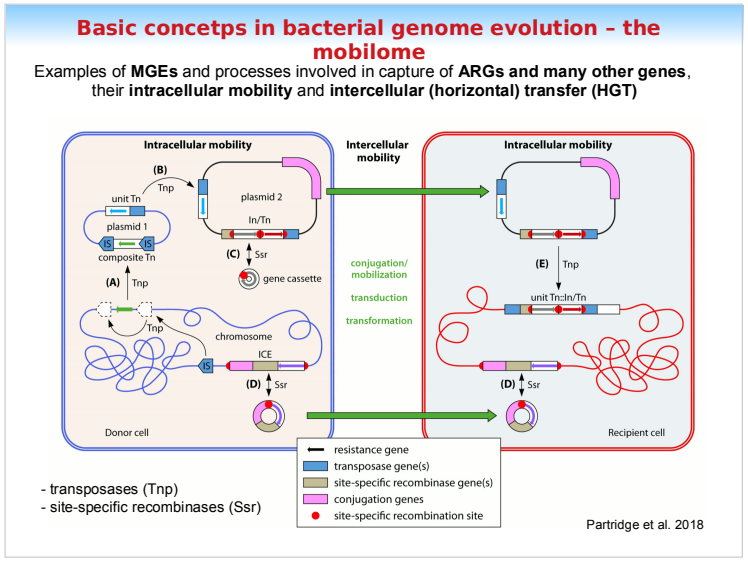
Syngamic sex vs. prokaryotic parasexuality:

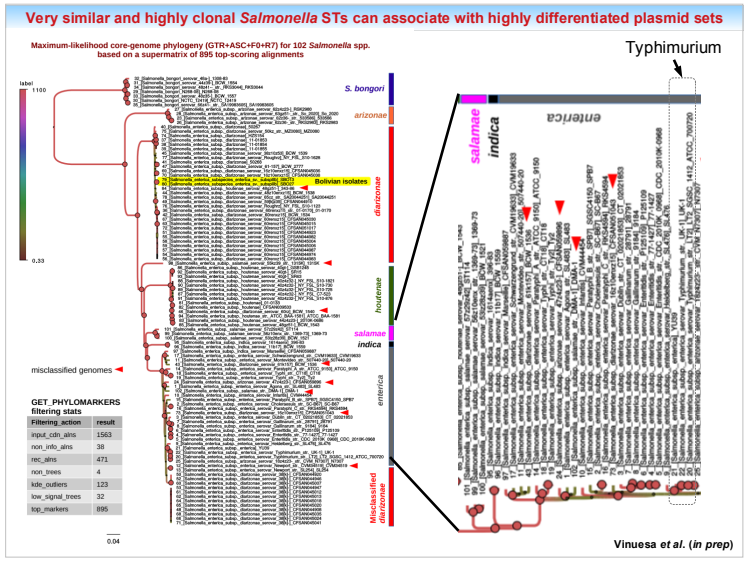
i) Bacterial sex is rare and decoupled from reproduction
ii) Can be highly promiscuous
iii) It is highly localized and mediated by different homologous and illegitimate recombination mechanisms

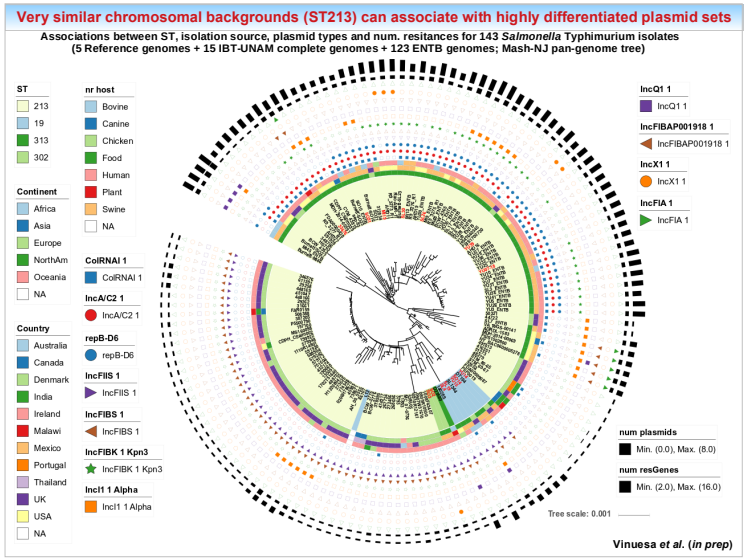
1. **homologous recomb. (*recA*)**
2. **illegitimate recomb. (transposases integrases)**

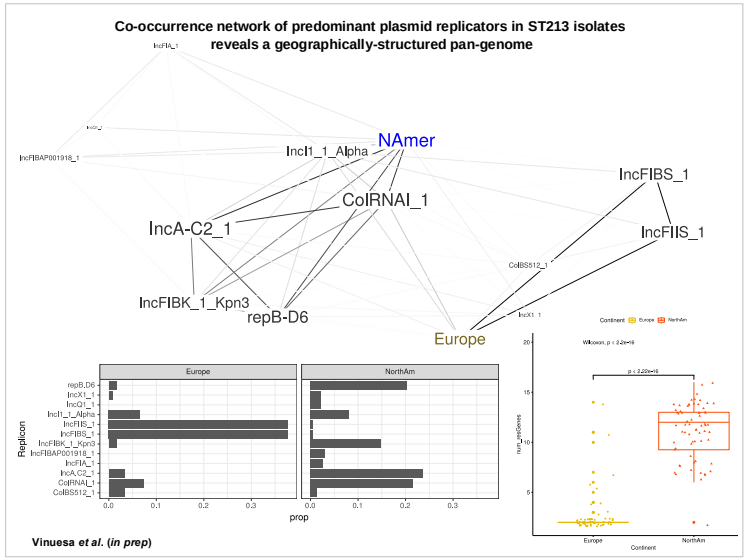
Mobile genetic elements and the accessory genome

- **Insertion sequences and compound transposons**
- **Integrons and gene cassettes**
- **Mobile and mobilizable genomic islands (GIs)**
  - **ICE: integrative and conjugative elements**
- **Prophages**
- **Plasmids**

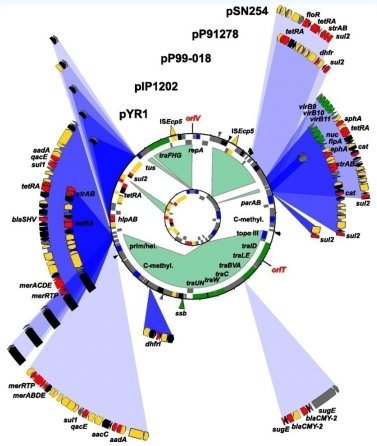


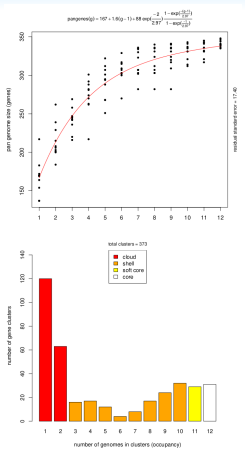






Large plasmids and other MGEs frequently have a complex mosaic structure and open pan-genome - The case of conjugative multi-resistance IncC plasmids

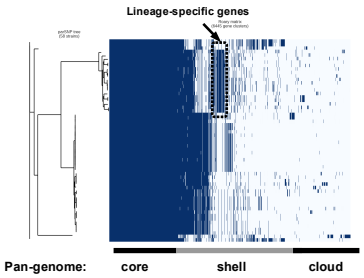


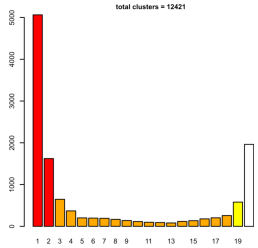


Fricke W F et al. J. Bacteriol. 2009;191:4750-4757

Vinuesa & Contreras-Moreira, 2015

Computational methods in pan-genome and phylogenomic research

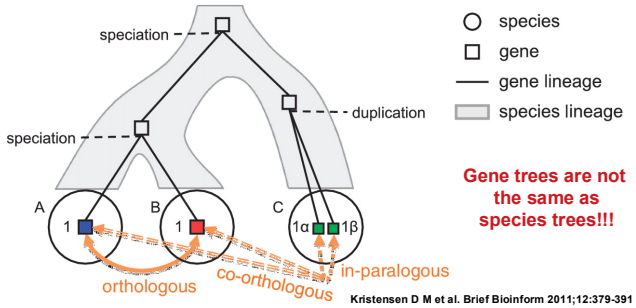




- **Inferring clusters of homologous sequences from mosaic genomes**
- **Computing and modeling pan-genomes**
- **Inferring genome phylogenies from core- and pan-genome data**

Basic concepts in bacterial genome evolution - homology relationships between gene families

Orthology, co-orthology and paralogy relationships in the evolution of four genes that arose from a single common ancestor.



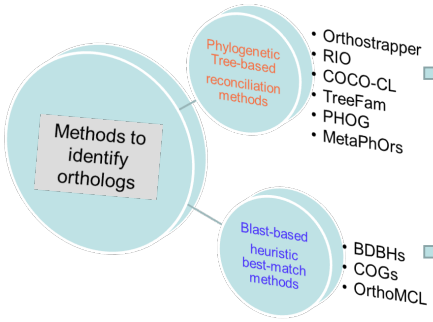
Gene trees are not the same as species trees!!!

Kristensen D M et al. Brief Bioinform 2011;12:379-391

Orthologs: essentialy instances of 'the same gene' in different species. Tend to retain function across large phylogenetic distances. Key markers for phylogenetics and genome annotation.

Paralogs: Tend to diverge over time to perform different functions via subfunctionalization or neofunctionalization routes.

Computational methods to identify homologs



- Most direct approach but known to be subjected to increasing artifacts as evol. dist. increases due to:
  - alignment problems
  - long-branch attraction
  - heterotachy
  - tree searches stuck in local sub-optimal maxima
- More adequate for bacteria, as they don't follow a clear tree-like evolutionary path (HGT).
- The only practical strategy for large datasets, therefore most common method in comparative genomics

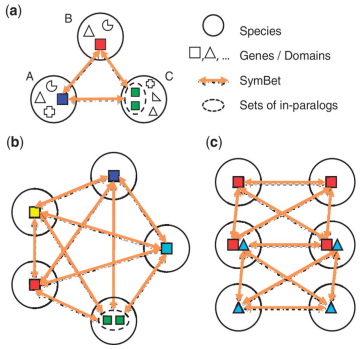
Computational methods to identify homologs

Grouping of genes in different species that are each others' BLAST BBHs into sets of orthologs and co-orthologs.

Linking pairs of BBHs from multiple genomes has a property of self-verification, as their consistency would be very unlikely due to chance, especially between phylogenetically distant lineages.

Methods for clustering of pairwise BBHs vary, but the most widely used approach involves a **single-linkage clustering procedure, where any two clusters sharing a common BBH are merged until convergence**.

Problems with: differential gene loss, domain recombination/gain/loss



Kristensen D M et al. Brief Bioinform 2011;12:379-391

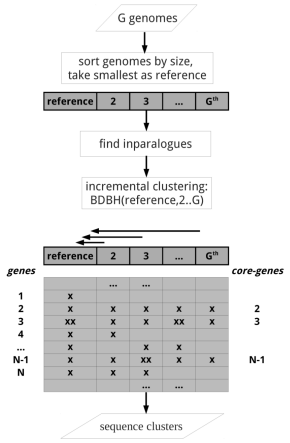
Computational methods to identify homologs: BDBHs

The bidirectional best-hit approach (BDBH)

As implemented in **GET_HOMOLOGUES**

(Contreras-Moreira & Vinuesa 2013)

1. Find inparalogues in all genomes
2. Find reciprocal best blast hits between reference and remaining genomes
3. Cluster sequences based on filtering criteria such as:
   - % alignment coverage
   - min. % sequence identity
   - E-score cut-off
   - Pfam domain-composition
   - sinteny

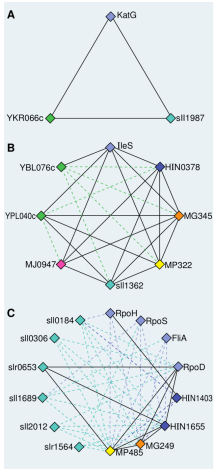


Computational methods to identify homologs: COGtriangles

Method	Description
Clusters of orthologous groups (COGs), variants and derivatives	Identifies **three-way BBHs** between orthologs or sets of co-orthologs **in three different species**, and these groups expanded (merging triangles whenever they share a common side) until saturation, followed by manual splitting of large groups improperly joined by multidomain proteins or complex mixtures of in- and out-paralogs.

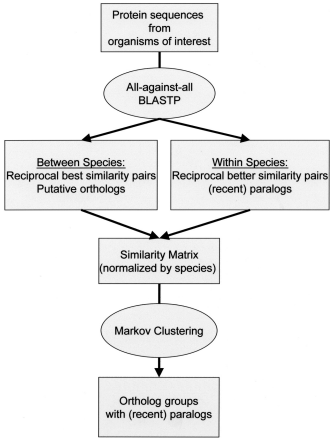
Tatusov et al. Science 1997. Vol. 278:631-637

Kristensen et al. Bioinformatics 2010. Vol. 26:1481-1487



Computational methods to identify homologs: OrthoMCL

Method	Description
OrthoMCL	Forms groups of orthologs and co-orthologs using a **Markov clustering process** involving iterative simulations of **stochastic (randomized) flow on the edges of a BBH graph**, with clusters of desired tightness identified depending on a given 'inflation' parameter determined by trial and error.



Li L et al. Genome Res. 2003;13:2178-2189

Open-source tools for phylogenomics and microbial pan-genomics
GET HOMOLOGUES & GET PHYLOMARKERS

Pablo Vinuesa
vinuesa

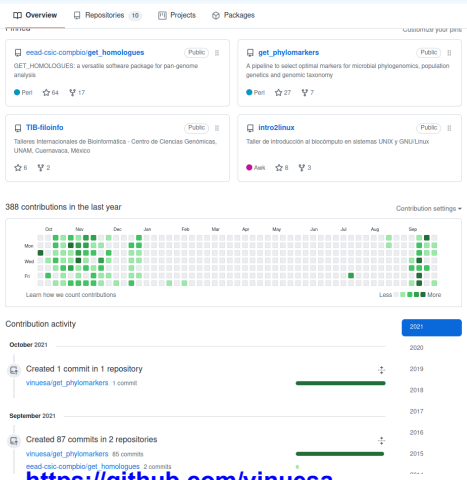
I'm a microbiologist interested in genomics, ecology and evolution. I develop open-source code for these topics in collaboration with @eead-csic-compbio

Edit profile

46 followers · 19 following · 10

@CCG-UNAM
Cuernavaca, Morelos, Mexico
http://www.ccg.unam.mx/~vinuesa/

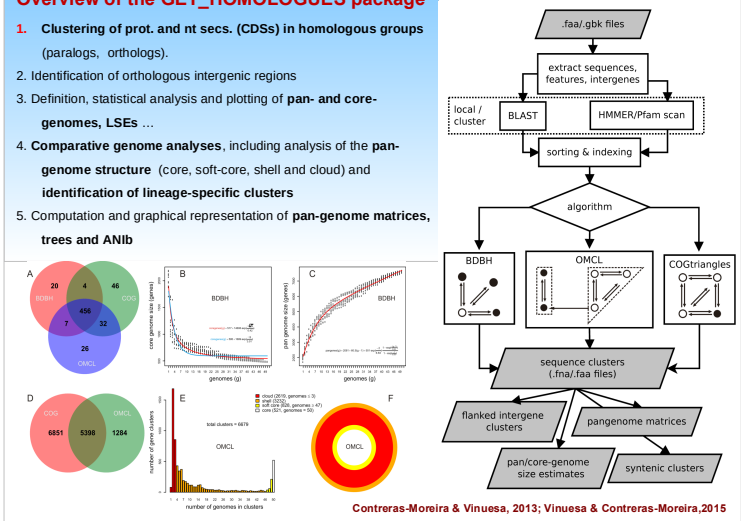
Achievements



https://github.com/vinuesa

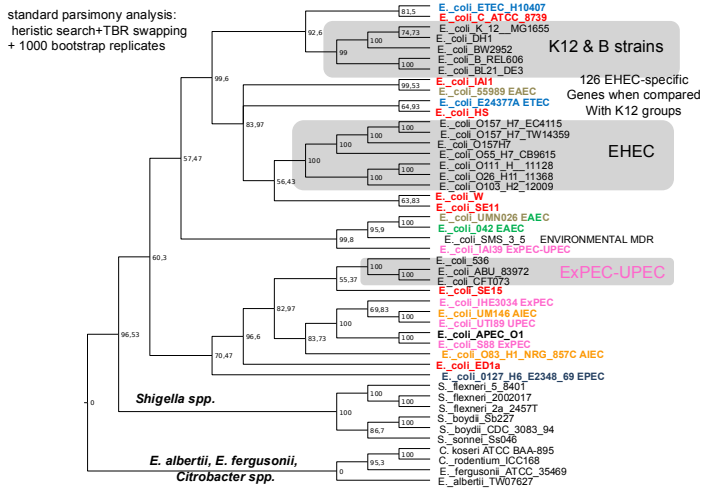
Overview of the GET_HOMOLOGUES package

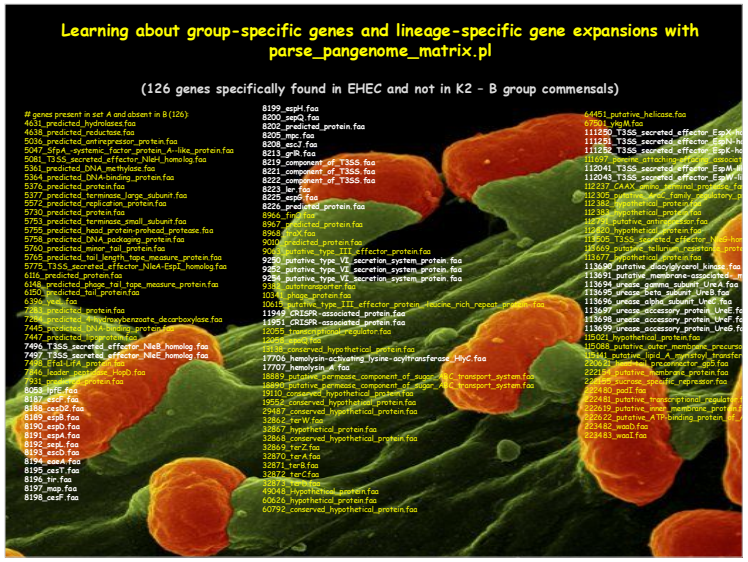
1. **Clustering of prot. and nt secs. (CDSs) in homologous groups** (paralogs, orthologs).
2. Identification of orthologous intergenic regions
3. Definition, statistical analysis and plotting of **pan- and core-genomes, LSEs** ...
4. **Comparative genome analyses**, including analysis of the **pan-genome structure** (core, soft-core, shell and cloud) and **identification of lineage-specific clusters**
5. Computation and graphical representation of **pan-genome matrices, trees and ANIb**

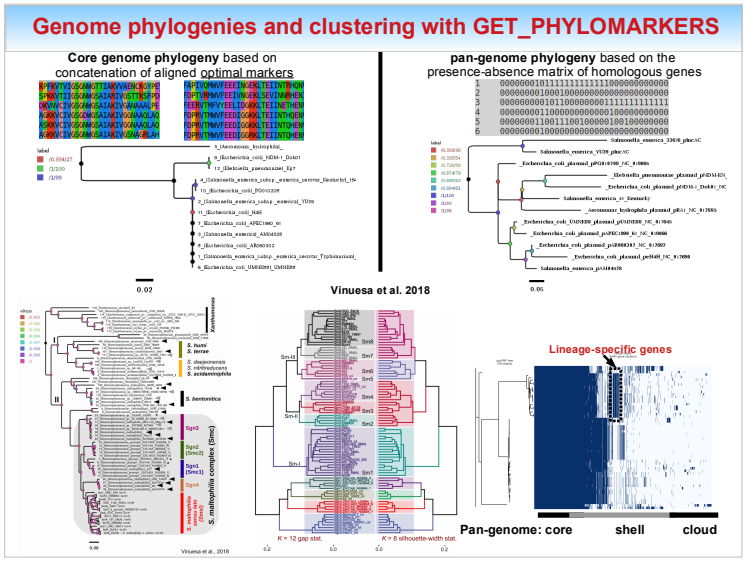


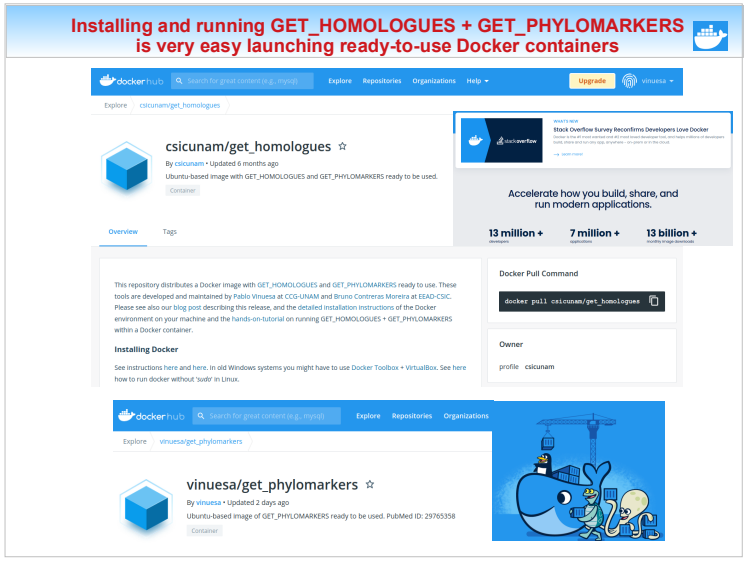
Contreras-Moreira & Vinuesa, 2013; Vinuesa & Contreras-Moreira,2015

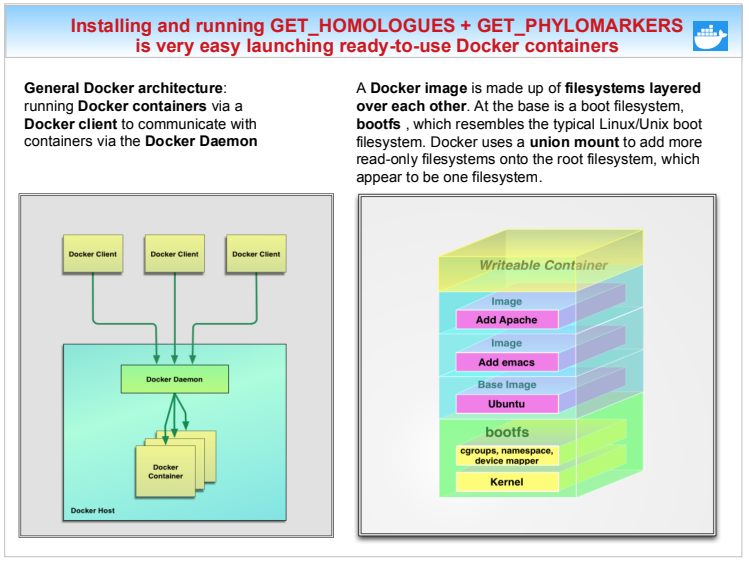
E. coli pangenome phylogeny esimated using the consensus panmatrix of presence/absence data for 46 fully sequenced genomes of *Escherichia* spp. / *Shigella* spp. and *Citrobacter* spp.











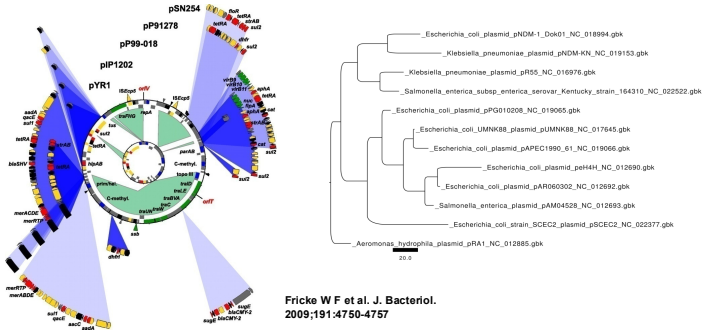
Basic Docker commands and use of containers

```
## To avoid permission errors (and the use of sudo), add your user to the docker group
# https://docs.docker.com/install/linux/linux-postinstall/
sudo groupadd docker
sudo usermod -aG docker $USER

# 1. Get general docker info and print help
$ docker info
$ docker --help
$ docker run --help

# 2. List available docker images on your system
$ docker image ls

# 3. List running containers
$ docker container ls

# 4. Stop a container
$ docker container stop CONTAINER-ID

# 5. Pull a Docker image from the registry
$ docker pull csicunam/get_homologues:latest

# 6. Launch an image, using a mount-bind of a user directory on the Docker container
$ docker run --rm -it -v $HOME/get_homPhy:/home/you/get_homPhy \
   csicunam/get_homologues:latest /bin/bash

# The last command uses options --rm to remove the container after exiting and sets an
-i interactive session calling a pseudo tty (-t), mounting a host directory on the
container (-v ...), accessible for wr from both, and launching a bash shell (/bin/bash)
```

Robust identification of orthologues and paralogues for microbial pan-genomics using GET_HOMOLOGUES: a case study of pIncA/C plasmids

Pablo Vinuesa[1] and Bruno Contreras-Moreira[1,2,3]
1 Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico.
2 Fundación ARAID, Zaragoza, Spain.
3 Estación Experimental de Aula Dei, Consejo Superior de Investigaciones Cientificas (EEAD-CSIC). Avda. Montañana, 1005, 50059 Zaragoza, Spain.

In: "Bacterial Pan-genomics". Methods in Molecular Biology. Marco Galardini, Alessio Mengoni and Marco Fondi Eds. Humana Press, Springer, 2014. *In press*



Fricke W F et al. J. Bacteriol. 2009;191:4750-4757

The GET_HOMOLOGUES + GET_PHYLOMARKERS tutorials:

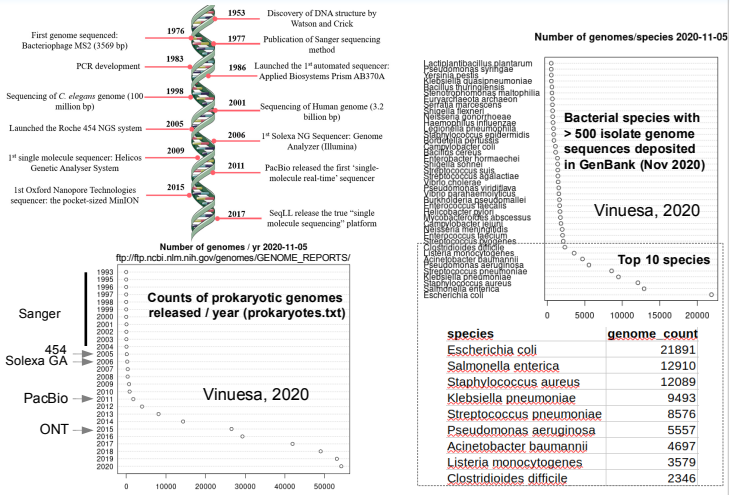
- https://github.com/vinuesa/get_phylomarkers/

Analyses to be performed in an upcoming practical session:

A pangenomic analysis of pIncA/C plasmids using GET_HOMOLOGUES

- Defining a robust core- and pan-genome of pIncA/C plasmids
- Exploring the gene space of pIncA/C plasmids: core, shell, cloud
- Pan-genome trees vs. core genome trees (supermatrices)
- Identifying lineage-specific genes in NDM-1 producing plasmids

Genome sequencing and the challenge of large-scale pan-genome analysis



species	genome_count
Escherichia coli	21891
Salmonella enterica	12910
Staphylococcus aureus	12089
Klebsiella pneumoniae	9493
Streptococcus pneumoniae	8576
Pseudomonas aeruginosa	5557
Acinetobacter baumannii	4697
Listeria monocytogenes	3579
Clostridioides difficile	2346

Genome sequencing and the challenge of large-scale pan-genome analysis

Roary: rapid large-scale prokaryote pan genome analysis

Bioinformatics, 31(22), 2015, 3691–3693

Andrew J. Page[1,*], Carla A. Cummins[1], Martin Hunt[1], Vanessa K. Wong[1,2], Sandra Reuter[2], Matthew T.G. Holden[3], Maria Fookes[1], Daniel Falush[4], Jacqueline A. Keane[1] and Julian Parkhill[1]

- Standard pan-genome software approaches

- The construction of a pan genome is **NP-hard** (Nguyen et al. 2014) with additional difficulties from real data due to contamination, fragmented assemblies and poor annotation. Therefore, any approach requires **heuristics** to produce a pan genome
- **all-against-all** comparison using **BLAST**, with the **running time growing ~quadratically** with the size of input data and are computationally infeasible with large datasets.
- They also have **quadratic memory requirements**, quickly exceeding the RAM available in high performance servers for large datasets.

- Roary's heuristic approach

- Roary address the computational issues by performing a **rapid clustering of highly similar sequences with CD-HIT** which can reduce the running time of BLAST substantially, and carefully manages RAM usage so that it increases linearly, both of which make it possible to analyze datasets with thousands of samples using commonly available computing hardware.

Genome sequencing and the challenge of large-scale pan-genome analysis

- Roary's pipeline (Page et al. (2015) Bioinformatics, 31(22), 2015, 3691–3693

1. Prokka-annotated genomes (**GFF3 + FASTA**)
2. Extract all the coding sequences, convert them into protein, and **cluster with CD-HIT**
3. An **all-against-all** comparison is performed with **BLASTP** on the clustered sequences with a user defined percentage sequence identity (default 95%)
4. Sequences are then clustered with **MCL** (Enright et al., 2002)
5. Finally, the pre-clustering results from CD-HIT are merged together with the results of MCL
6. Using conserved gene neighborhood information, homologous groups containing paralogs are split into groups of true orthologs
7. A **graph** is constructed **of the relationships of the clusters based on the order of occu-rrence in the input sequences**, allowing for the clusters to be ordered and thus **providing context for each gene**.
8. Isolates are clustered based on gene presence in the accessory genome, with the contribution of isolates to the graph weighted by cluster size.
9. A suite of command line tools is provided to interrogate the dataset providing union, intersection and complement.

Genome sequencing and the challenge of large-scale pan-genome analysis

- Roary's pipeline benchmark (Page et al. (2015) Bioinformatics, 31(22), 2015, 3691–3693

- 1 processor (AMD Opteron 6272), 60 GB of RAM.
- Constructed a simulated dataset based on Salmonella Typhi (S.typhi) CT18, to accurately assess the quality of the clustering.
- Created 12 genomes with 994 identical core genes and 23 accessory genes in varying combinations
- The running time and RAM of PGAP and PanOCT increases substantially with dataset size; LS-BSR, over clusters in 2% of cases.

Table 1. Accuracy of each pan genome application on a dataset of simulated data

	Core genes	Total genes	Incorrect split	Incorrect merge
Expected	994	1017	0	0
PGAP	991	1012	0	4
PanOCT	993	1015	1	1
LS-BSR	974	994	0	23
Roary	994	1017	0	0

Table 2. Comparison of pan genome applications using real *S.typhi* data (ERP001718)

Samples	Software	Core*	Total	RAM (mb)	Wall time (s)
8	PGAP	4545	4929	569	41 397
	PanOCT	4544	4936	663	1457
	LS-BSR	4476	4816	270	2585
	Roary	4459	4871	156	44
24	PGAP	—	—	—	—
	PanOCT	4522	4991	5313	96 093
	LS-BSR	4451	4843	554	7807
	Roary	4436	4941	444	382
1000	PGAP	—	—	—	—
	PanOCT	—	—	—	—
	LS-BSR	4272	7265	17 413	345 019
	Roary	4016	9201	13 752	15 463

*Core is defined as a gene being in at least 99% of samples, which allows for some assembly errors in very large datasets. Where there are no results, the applications failed to complete within 5 days or used more than 60 GB of RAM. The first column is the number of unique *S.typhi* genomes in the input set with a mean of 54 contigs over all 1000 assemblies.

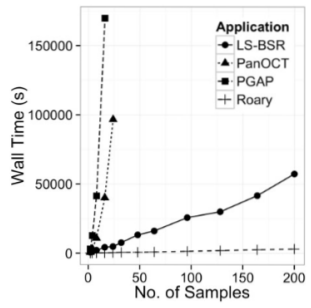


Fig. 1. Effect of dataset size on the wall time of multiple applications. Onl analysis that completed within 2 days and 60 GB of RAM is shown

Taking population-structure into account subdivides the pan-genome into 13 categories

MICROBIAL GENOMICS

RESEARCH ARTICLE

Horesh *et al.*, *Microbial Genomics* 2021;7:000670
DOI 10.1099/mgen.0.000670

Different evolutionary trends form the twilight zone of the bacterial pan-genome

Gal Horesh[1], Alyce Taylor-Brown[1], Stephanie McGimpsey[1], Florent Lassalle[1], Jukka Corander[1,2,3], Eva Heinz[4,*] and Nicholas R. Thomson[1,5,*]

Abstract

The pan-genome is defined as the combined set of all genes in the gene pool of a species. Pan-genome analyses have been very useful in helping to understand different evolutionary dynamics of bacterial species: an open pan-genome often indicates a free-living lifestyle with metabolic versatility, while closed pan-genomes are linked to host-restricted, ecologically specialized bacteria. A detailed understanding of the species pan-genome has also been instrumental in tracking the phylodynamics of emerging drug resistance mechanisms and drug-resistant pathogens. However, current approaches to analyse a species' pan-genome do not take the species population structure into account, nor do they account for the uneven sampling of different lineages, as is commonplace due to over-sampling of clinically relevant representatives. Here we present the application of a population structure-aware approach for classifying genes in a pan-genome based on within-species distribution. We demonstrate our approach on a collection of 7500 *Escherichia coli* genomes, one of the most-studied bacterial species and used as a model for an open pan-genome. We reveal clearly distinct groups of genes, clustered by different underlying evolutionary dynamics, and provide a more biologically informed and accurate description of the species' pan-genome.

The pan-genome has a heterogeneous distribution across the specie's members. Traditionally, this gene pool has been divided into **core genes**, present across the majority of genomes, and **accessory genes**, whose presence vary across the dataset. These traditional methods **do not reflect the true complexity of gene dynamics across a diverse species**, nor do they account for **population structure or the inherent biases in sampling and sequencing.**

Horesh et al. propose a novel framework that further divides the core and accessory categories into 13 subcategories to better account for differences between lineages at a finer scale than traditional methods.

Taking population-structure into account subdivides the pan-genome into 13 categories

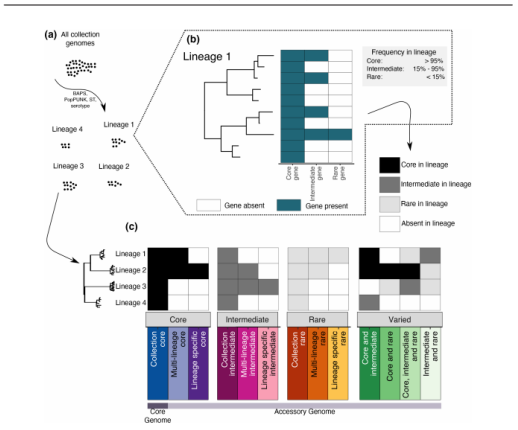


Fig. 1. Twilight pan-genome analysis workflow. (a) A collection of genomes are grouped into lineages of closely related isolates. (b) Each gene is classified as core, intermediate or rare in each lineage, depending on its frequency within the lineage (as defined in the grey box). (c) The classification of the entire gene pool across all lineages consists of a total of 13 distribution classes. These include the number of lineages in which a gene is present (all lineages, multiple lineages or a single lineage), and the combination of frequency assignments of the gene in those lineages (core, intermediate or rare).

Why do prokaryotes have pan-genomes?

Why prokaryotes have pangenomes

James O. McInerney[1*], Alan McNally[2] and Mary J. O'Connell[3]

The existence of large amounts of within-species genome content variability is puzzling. Population genetics tells us that fitness effects of new variants—either deleterious, neutral or advantageous—combined with the long-term effective population size of the species determines the likelihood of a new variant being removed, spreading to fixation or remaining polymorphic. Consequently, we expect that selection and drift will reduce genetic variation, which makes large amounts of gene content variation in some species so puzzling. Here, we amalgamate population genetic theory with models of horizontal gene transfer and assert that pangenomes most easily arise in organisms with large long-term effective population sizes, as a consequence of acquiring advantageous genes, and that the focal species has the ability to migrate to new niches. Therefore, we suggest that pangenomes are the result of adaptive, not neutral, evolution.

Nat. Microbiol. 28 MARCH 2017 | VOLUME: 2 | 17040

- **There is a controversy around the question: are pangenomes adaptive?**
  - HGT-mediated introgression of advantageous genes could lead to a **selective sweep**, reducing diversity
  - This would happen particularly in **populations with small Ne** (num. of individuals that contribute offspring to the next generation)

- **An adaptive model for HGT when coupled with migration**
  - **A new compartment model by Niehus et al. that explicitly models HGT and migration has shown the plausibility of selection on HGT genes driving population differentiation**. Using a mathematical approach, the authors showed that in the case of a selectively advantageous HGT event, diversity is removed from the species when there is no migration into or out of the compartment or niche occupied by the focal prokaryotic community.

  - By contrast, a model that includes migration to and from the niche, combined with HGT of a selectively advantageous gene, can theoretically result in a situation where diversity is not necessarily reduced.

- There is a growing body of evidence that accessory genes might provide significant benefit. Karcagi et al. analysed a range of *E. coli* genomes at different levels of gene deletion, specifically genes that had been recently acquired by HGT. They found that HGT genes conferred significant benefits in terms of substrate utilization, efficiency of resource usage to build new cells, and tolerance to stress. Loss of HGT genes tended to affect fitness in several measurable ways, including the induction of a general stress response, inability to grow at all in some environments …

Why do prokaryotes have pan-genomes?

Why prokaryotes have pangenomes

James O. McInerney[1*], Alan McNally[2] and Mary J. O'Connell[3]

The existence of large amounts of within-species genome content variability is puzzling. Population genetics tells us that fitness effects of new variants—either deleterious, neutral or advantageous—combined with the long-term effective population size of the species determines the likelihood of a new variant being removed, spreading to fixation or remaining polymorphic. Consequently, we expect that selection and drift will reduce genetic variation, which makes large amounts of gene content variation in some species so puzzling. Here, we amalgamate population genetic theory with models of horizontal gene transfer and assert that pangenomes most easily arise in organisms with large long-term effective population sizes, as a consequence of acquiring advantageous genes, and that the focal species has the ability to migrate to new niches. Therefore, we suggest that pangenomes are the result of adaptive, not neutral, evolution.

Nat. Microbiol. 28 MARCH 2017 | VOLUME: 2 | 17040

Conclusion

- The authors infer that effective population size and the ability to migrate to new niches are the most influential factors in determining pangenome size.

- Given the huge size of pan-genome estimates implies that the number of ecological niches on the planet must be enormous, which is in line with recent analysis of genomic diversity that suggests there may be 10E12 (one trillion) microbial species on Earth.

- That the majority of genes in the biosphere are not strongly attached to any group of organisms has been a surprise of the genomics era, and consequently this 'public goods' hypothesis needed explanation

Bacterial evolution, mobile genetic elements and the pan-genome

References

*** Homology and genome evolution**

Fitch WM. Homology a personal view on some of the problems. Trends Genet. 2000 May;16(5):227-31

Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 2005;39:309-38

Koonin EV, Wolf YI. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res. 2008 Dec;36(21):6688-719.

Kristensen DM, Wolf YI, Mushegian AR, Koonin EV. Computational methods for Gene Orthology inference. Brief Bioinform. 2011 Sep;12(5):379-91. **<=**

*** Bacterial mobile genetic elements**

Aziz RK, Breitbart M, Edwards RA. Transposases are the most abundant, most ubiquitous genes in nature. Nucleic Acids Res. 2010 Jul;38(13):4207-17

Bellanger X, Payot S, Leblond-Bourget N, Guédon G. Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. FEMS Microbiol Rev. 2014 Jul;38(4):720-60

Cambray G, Guerout AM, Mazel D. Integrons. Annu Rev Genet. 2010;44:141-66

Campbell A. The future of bacteriophage biology. Nat Rev Genet. 2003 Jun;4(6):471-7.

Frost LS, Leplae R, Summers AO, Toussaint A. Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol. 2005 Sep;3(9):722-32

Galán JE, Waksman G. Protein-Injection Machines in Bacteria. Cell. 2018 Mar 8;172(6):1306-1318

Gillings MR. Integrons: past, present, and future. Microbiol Mol Biol Rev.2014 Jun;78(2):257-77

Johnson CM, Grossman AD. Integrative and Conjugative Elements (ICEs): What They Do and How They Work. Annu Rev Genet. 2015;49:577-601

Juhas M, van der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW. Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol Rev. 2009

Partridge SR, Kwong SM, Firth N, Jensen SO. Mobile Genetic Elements Associated with Antimicrobial Resistance. Clin Microbiol Rev. 2018 Aug 1;31(4) **<=**

Pilla G, Tang CM. Going around in circles: virulence plasmids in enteric pathogens. Nat Rev Microbiol. 2018 Aug;16(8):484-495 **<=**

Bacterial evolution, mobile genetic elements and the pan-genome

References

*** Bacterial mobile genetic elements (continuation)**

Siguier P, Gourbeyre E, Varani A, Ton-Hoang B, Chandler M. Everyman's Guide to Bacterial Insertion Sequences. Microbiol Spectr. 2015 Apr;3(2):MDNA3-0030-2014

Welch RA, Burland V, Plunkett G 3rd, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HL, Donnenberg MS, Blattner FR. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc Natl Acad Sci U S A. 2002 Dec 24;99(26):17020-4

Wozniak RA, Waldor MK. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. Nat Rev Microbiol. 2010 Aug;8(8):552-63.

Bacterial evolution, mobile genetic elements and the pan-genome

References

*** Mosaic genome structure, population structure and the bacterial pan-genome: reviews and papers**

Golicz, A. A., Bayer, P. E., Bhalla, P. L., Batley, J. & Edwards, D. Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications. Trends in Genetics vol. 36 132–145 (2020). **<=**

Horesh et al. 2021. Different evolutionary trends form the twilight zone of the bacterial pan-genome. Microb Genom. 2021 Sep;7(9). **<=**

McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. Nat Microbiol. 2017 Mar 28;2:17040 **<=**

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Curr Opin Genet Dev. 2005 Dec;15(6):589-94

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A. 2005 Sep 27;102(39):13950-5. doi: 10.1073/pnas.0506758102. Epub 2005 Sep 19. Erratum in: Proc Natl Acad Sci U S A. 2005 Nov 8;102(45):16530. PMID: 16172379; PMCID: PMC1216834.

Tettelin H, Medini D, editors. The Pangenome: Diversity, Dynamics and Evolution of Genomes. Cham (CH): Springer; 2020. PMID: 32633920.

Welch RA, Burland V, Plunkett G 3rd, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HL, Donnenberg MS, Blattner FR. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc Natl Acad Sci U S A. 2002 Dec 24;99(26):17020-4

*** Tools and approaches for computing pan-genomes and (pan-)genome phylogenies**

Contreras-Moreira B, Vinuesa P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol. 2013 Dec;79(24):7696-701

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015 Nov 15;31(22):3691-3.

Vinuesa P, Contreras-Moreira B. Robust identification of orthologues and paralogues for microbial pan-genomics using GET_HOMOLOGUES: a case study of pIncA/C plasmids. Methods Mol Biol. 2015;1231:203-32

Vinuesa P, Ochoa-Sánchez LE, Contreras-Moreira B. GET_PHYLOMARKERS, a Software Package to Select Optimal Orthologous Clusters for Phylogenomics and Inferring Pan-Genome Phylogenies, Used for a Critical Geno-Taxonomic Revision of the Genus Stenotrophomonas. Front Microbiol. 2018 May 1;9:771